Hotel Bookings Exploratory Data Analysis

Tushar Gautam
Data science trainees,
AlmaBetter, Delhi

Abstract:

When we hear about hotel booking, we definitely hear about Trivago. Thanks to large scale advertisement, I can't seem to get ad out of my head. Whenever someone try to book hotel, they will surely consider following factors like price per night, distance of hotel, restaurant availability, scenery location, room quality, cleanliness, food quality.

In this dataset we are provided with dataset that contains data about hotel bookings, cancellation and some important factors that relates to hotel bookings. Our analysis will help us to understand what factors are important for hotel booking, preference of guest when they book hotels etc. We will analyze hotel bookings data to gain insights.

Keywords: Hotel bookings, City hotels, Resort hotels, Cancellation, Distribution channel.

1.Problem Statement:

We are provided with data of hotel bookings we had analyze data on following questions.

- Which Hotel has the most number of bookings?
- Which year had the maximum number of bookings throughout and on hotel basis?
- Which Month had the Maximum Number of Bookings?
- Which Month had the Maximum Number of Bookings in each Year?
- Which Week had the Maximum number of bookings throughout as well as in the year with Maximum bookings? i.e Year - 2016
- What type of meals corporate prefer during there visit?
- Which hotel had most number of cancellation Throughout?
- Which year has the highest number of cancellations Hotel-Wise?
- Which is the Most Reserved Room Type?
- Which Countries had most number of Bookings?

2.Data description:

The main objective of Exploratory data analysis is to understand trend and behavior of guest in hotel bookings. For that first we will need to understand what every feature in data means.

The data table consists of 119,390 rows and 32 columns. Our analysis starts with defining each column and our understanding for each column mentioned below:

- hotel: Hotel type (City hotels, Resort hotels)
- is_canceled: value indicates if the booking is cancelled or not.
- lead_time: How long in advance the booking was made.

- arrival_date_year: Customer arrival year.
- arrival_date_month: In which month of the year customer visited hotel.
- arrival_date_week_number: In which week of the year customer arrived.
- arrival_date_day_of_month: Date of the month customer visited hotel.
- stays_in_weekend_nights: Customer stayed or booked to stay in hotel during weekend nights.
- stays_in_week_nights: Customer stayed in hotel during week nights.
- adults: Number of adults.
- children: number of children.
- babies: Number of babies.
- meal: Type of meal booked.:
- country: Country of orgin of cutomer.
- market_segment: where the bookings came from.
- distribution_channel: Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
- is_repeated_guest: Value indicating if the booking name was from a repeated guest (1) or not (0).
- previous_cancellations: Number of previous bookings that were cancelled by the customer prior to the current booking.
- previous_bookings_not_canceled: umber of previous bookings that were cancelled by the customer prior to the current booking.
- reserved_room_type: Code of room type reserved. Code is presented instead of designation for anonymity reasons.
- assigned_room_type: Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due.
- booking_changes: Number of changes/amendments made to the booking from the moment the booking was entered on the PMS.
- deposit_type: Indication on if the customer made a deposit to guarantee the booking.
- agent: ID of the travel agency that made the booking.
- company: ID of the company/entity that made the booking or responsible for paying the booking.
- days_in_waiting_list: Number of days the booking was in the waiting list before it was confirmed to the customer.
- customer_type: Type of booking, assuming one of four categories.
- adr: Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights.
- required_car_parking_spaces: Number of car parking spaces required by the customer.
- total_of_special_requests: Number of special requests made by the customer (e.g. twin bed or high floor).
- reservation_status: Reservation last status, assuming one of three categories: Canceled –booking was canceled by the customer; CheckOut: customer check out

from hotel,Noshow: Customer did not check-in hotel and informed hotel with reason.

- reservation_status_date: Date at which the last status was set. This variable can be used in conjunction with the Reservation Status to understand when was the booking cancelled or when did the customer checked out of the hotel.

3.Introduction:

This dataset with hotel bookings data. One of the hotels is a resort hotel and the other is a city hotel. The dataset consists of 32 columns and 119,390 observations. Each observation represents a hotel booking. The datasets comprehend hotel bookings to arrive between the year 2015 and the 2017, including bookings that effectively arrived and bookings that were canceled. Due to the scarcity of real business data for scientific and educational purposes, these datasets can have an important role for research and education in revenue management, machine learning, or data mining, as well as in other fields.

Our analysis will help us to understand data in depth, trends in hotel bookings that can help us to plan better deal accordingly for both hotels and guests in future.

4. How hotel reservation system works?

A **hotel reservation system** is the mechanism through which guests can create secure online reservations. While the process is similar to booking with an online travel agent (OTA), the difference is the hotel's booking engine essentially links up to their own website so that there are no additional fees incurred for the property.

Through the hotel reservation system software, guests can choose how long they will stay, the type of room they want, get add-ons, and pay securely online through a payment platform. In this article we'll go in depth about what a hotel reservation system is, how it works, and the benefits that it offers to the hospitality sector. So, let's get started.

A hotel reservation system is a software application that allows guests to book directly with the hotel online, with no intermediaries necessary. The software essentially processes online reservations made via the hotel's website and then passes this information to the hotel's own backend so that the information can be easily accessed. Bookings are then managed by hotel staff.

With the boom of the Millennial traveler, now more than 700 million people are expected to book primarily online by 2023, so having an online reservation system is key to reaching a widespread audience. It is also key to generating a good first impression because guests are able to place bookings without having to navigate to

another domain. Keeping the whole process internally prevents clients from navigating away from the page before making the final booking.

5. Types of hotel bookings system?

There are mainly two type of hotel bookings systems Direct and Indirect.

5.1 Direct bookings

This booking source is important to most hotel owners and is considered the best long-term strategy for your hotel marketing and distribution strategy.

By offering direct booking, customers can make their room reservations by sending an email, calling the hotel, visiting and booking room services on your website, and social media channels.

Booking room through the hotel-owned website
Not only providing official information of hotel services, but the website is also a powerful tool to convert lookers to bookers without any cost. Therefore, it is very critical to build an online booking system for your site.

5.2 Indirect bookings

The current indirect booking channels are pretty diverse and popular with many customers.

The advantage of indirect booking channels is offering many choices for lookers. They can see many options from different hotels and compare prices, services, promotions, etc., to choose the right hotel for their needs.

Here are a few popular indirect booking channels.

OTA

OTA is an acronym for Online Travel Agency, this is a familiar room sales channel for those who are in the accommodation business.

Some OTA channels attracting a lot of users include Agoda, Traveloka, Booking, Expedia, Abay.

For each booking through the OTA channel, the hotel will have to pay a certain percentage of the booking value as commission for the room sales channel as agreed in the contract.

TA

TA stands for Travel Agency. These travel agencies often organize tours with a large and stable number of guests. Normally, travel agencies are responsible for arranging all services on tour including booking rooms for customers, which is why a TA will bring a stable number of customers to the hotel.

The guest source from TAs usually accounts for 50 – 60% of the hotel's bookings, so many hotels today are very focused on developing this indirect booking channel.

GDS

GDS stands for Global Distribution System, this is a global distribution system for hotel rooms, also known as the method of selling Business to business (B2B) rooms.

GDS works on the principle that hotels will sign a contract with GDS, then provide information about their offers. After that travel companies will get information from the GDS system and resell it to their tourists.

The hotel will have a lot of customers through the GDS.

6.Steps Involved

Exploratory Data Analysis

Exploratory Data Analysis is a data analytics process to understand the data in depth and learn the different data characteristics, often with visual means. This allows you to get a better feel of your data and find useful patterns in it

It is crucial to understand it in depth before you perform data analysis and run your data through an algorithm. You need to know the patterns in your data and determine which variables are important and which do not play a significant role in the output. Further, some variables may have correlations with other variables. You also need to recognize errors in your data.

All of this can be done with Exploratory Data Analysis. It helps you gather insights and make better sense of the data, and removes irregularities and unnecessary values from data.

6.1 Data Collection

Data collection is the process of collecting, measuring and analysing different types of information using a set of standard validated techniques. The main objective of data collection is to gather information-rich and reliable data, and analyse them to make critical business decisions. Once the data is collected, it goes through a rigorous process of data cleaning and data processing to make this data truly useful for businesses. It refers to the process of finding and loading data into our system.

Pandas library is used to loading our data in our system in python. Using pandas we can manipulate data easily.

6.2 Data Cleaning

Data cleaning refers to the process of removing unwanted variables and values from your dataset and getting rid of any irregularities in it. Such anomalies can disproportionately skew the data and hence adversely affect the results. Some steps that can be done to clean data are:

- Handling missing values: There are always some missing values in dataset. If we don't remove or handle those missing values then that can cause a trouble in our analysis. Removing or replacing those missing values with something meaningful is very important so that our data will have no missing values.
- Removing duplicates: Drop the duplicates rows.
- Formatting data to proper dtype.
- Adding or removing columns required for analysis.

6.3 Univariate or Bivariate analysis

Univariate analysis: In Univariate Analysis, you analyse data of just one variable. A variable in your dataset refers to a single feature/ column. You can do this either with graphical or non-graphical means by finding specific mathematical values in the data.

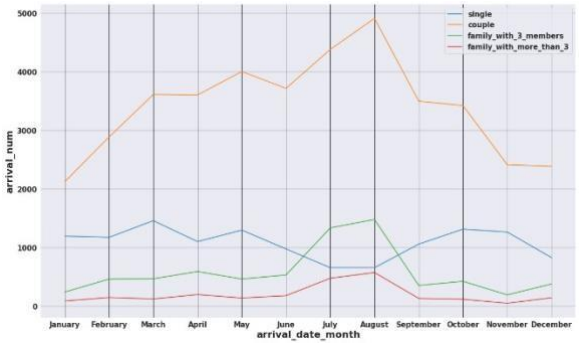
Bivariate analysis: Here, you use two variables and compare them. This way, you can find how one feature affects the other. It is done with scatter plots, which plot individual data points or correlation matrices that plot the correlation in hues. **6.4 Visualization**

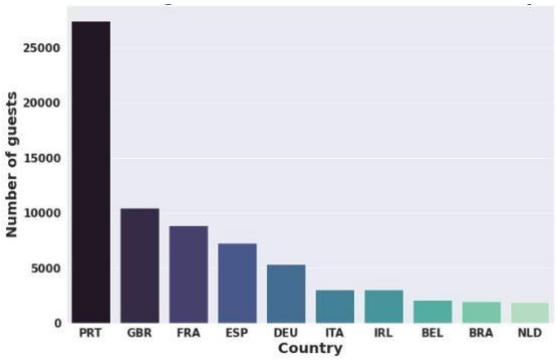
Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

Types of data visualizations

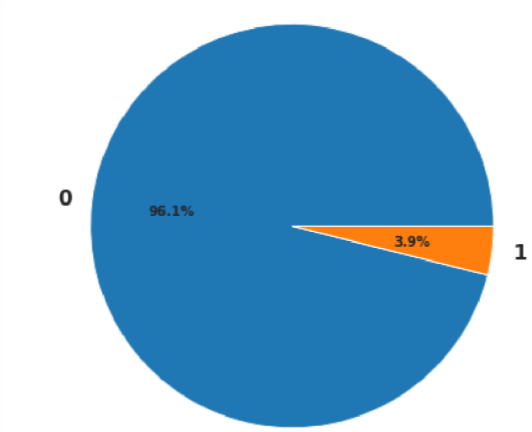
- Tables: This consists of rows and columns used to compare variables. Tables can show a great deal of information in a structured way, but they can also overwhelm users that are simply looking for high-level trends.

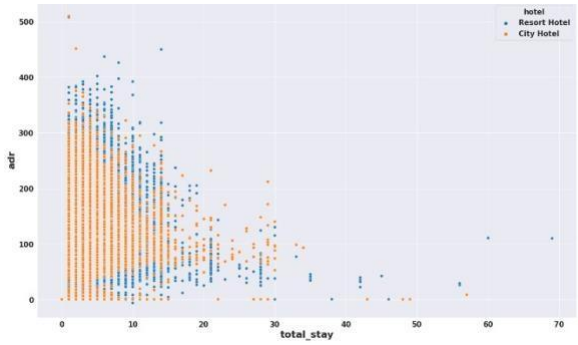
- Pie charts and bar charts: These graphs are divided into sections that represent parts of a whole. They provide a simple way to organize data and compare the size of each component to one other.





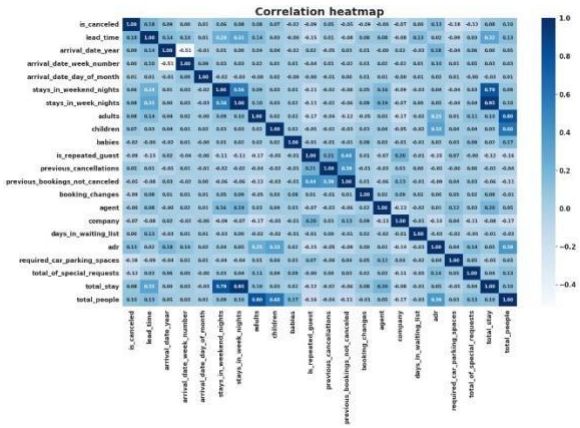
- Scatter plots: These visuals are beneficial in revealing the relationship between two variables, and they are commonly used within regression data analysis. However, these can sometimes be confused with bubble charts, which are used to visualize three variables via the x-axis, the y-axis, and the size of the bubble.





- Heat maps: These graphical displays are helpful in visualizing behavioural data by location. This visuals show change can be a location on a map, or even quantities by plotting a series of
- Line graphs and area charts: These in one or more a webpage.

data points over time. Line graphs utilize lines to demonstrate these changes while area charts connect data points with line segments, stacking variables on top of one another and using colour to distinguish between variables.



7. Conclusion

Now we reached to end of our project.

In all these processes we get to know that city hotels have more bookings than resort hotel might be because resort hotels are on costlier side or at longer distance on other has on other hand city hotels are less costly and situated near at Railway stations and Airports.

27 % of bookings were cancelled, only 3.9 % guests were repeated, guest preferred to stay for week or less than a week and Travel agents/Travel operators is the most preferred distribution channel for bookings.

References

[mews/hotel-reservation-system]

[hotellinksolutions Simplilearn]

[IBM data visualization]